Text Classification on Islamic Jurisprudence using Machine Learning Techniques

Text Classification on Islamic Jurisprudence using Machine Learning Techniques

K Jamal[1], R Kurniawan[2*], A S Batubara[3], M Z A Nazri[4], F Lestari[5] and P Papilo[6]

[1]Faculty of Usuluddin, Universitas Islam Negeri Sultan Syarif Kasim Riau, 28293Pekanbaru, Riau, Indonesia
[2] Department of Informatics Engineering, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Jl. HR. Subrantas Km. 15, Pekanbaru 28293, Indonesia
[3] Omdurman Islamic University, Sudan
[4] Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia
[5,6] Department of Industrial Engineering, Faculty of Science and Technology, Universitas Islam Negeri Sultan Syarif Kasim Riau, Jl. HR. Subrantas Km. 15, Pekanbaru 28293, Indonesia

*Corresponding e-mail: rahmadkurniawan@uin-suska.ac.id

Abstract. Indonesia, which is the world's most populous Muslim-majority nation, Islamic education is not an alternative to secular education but compulsory to any Muslims learn the way of life. The two primary sources of the religion of Islam is the Quran and Hadith. These two are where the majority of the teachings come from. However, when looking for guidance, a Muslim often refers to Islamic scholars to interpret the verses and hadith for them and educate themselves on a topic. However, due to many circumstances, people could not meet and ask a scholar personally regarding Islamic jurisprudence. Chatbot is one of the solutions. Chatbot offers users new opportunities to improve the learning and engagement process, reducing the typical cost of face-to-face consultation. However, to train a chatbot on Islamic Jurisprudence, a text classifier is needed to build a strong knowledge base for the chatbot. This study adopted a standard methodology for building a text classification model. This study used 600 Islamic jurisprudence text data obtained from the books and the web written by an influential Islamic scholar named Ustadz Abdul Somad (UAS). Machine learning algorithms such as Bayesian Network and Naïve Bayes, were employed to classify the text data. Based on the experimental testing results, the Naïve Bayes algorithm is more accurate in all evaluation models, 84.25% using training set, and 76.54% using 10-Fold cross-validation. Meanwhile, the Bayesian Network algorithm is faster in terms of time taken on all evaluation models. Thus, it can be concluded that the text classification model using Naïve Bayes and String to Word Vector filter have the potential to be used effectively but still has plenty of room for improvement.

1. Introduction

A chatbot or also known as virtual assistance is a computer program that is developed to replace human experts or human operators to communicate with human users through the internet. A chatbot is programmed to work independently without any supervision or assistance from a human. A human user usually did not notice that they are talking to a chatbot (or robot) until they posted complex

questions that are beyond the knowledge of the bot stored in its knowledge-base to answer the question.

Several chatbot or virtual assistants that are well known are Alexa, Cortana, or Siri. These chatbots use their wealth of knowledge to provide comprehensive answers to simple queries from users. Chatbot relies on Natural Language Processing (NLP) to convert a user's query (input text) into a structure that a computer can understand and then convert it into an internal query. Then, the chatbot obtains an output from the knowledge base, then deliver that output with a just text or text-to-speech exchange. The capability of a bot to communicate effectively is why they seem intelligent. However, without that knowledge base to draw upon, they are simple apps.

Virtual assistants can be one of the essential media for learning Islamic Jurisprudence because millions of Muslims would like to access to famous and knowledgeable scholars or experts. Islamic education is always on high demand as it is not just an alternative to secular education but also a way of life [1].

The sources of Fiqh are relevant, e.g., the *Qur'an*, *Hadith*, *Ijma* (collective thinking and agreement among a particular generation of authoritative Muslims and their interpretation by Islamic scholars), and *Qiyas* (an analogy is applied if there is no *Ijma* or historic collective reasoning published scientific).

Ustadz Abdul Somad is a busy Islamic Preacher, so that many people did not get assistance in the question and answer of Islamic jurisprudence. On the other hand, Ustadz Abdul Somad has published some books regarding questions and answers on Islamic jurisprudence. Nevertheless, ordinary people still have obstacles to understanding the many texts to get a conclusion. Therefore, artificial intelligence is considered as an innovative way to learn and understand Islamic jurisprudence easily.

Artificial intelligence (AI) technology is increasingly being exploited in various fields such as finance, military, medical, and even in Islamic[1,3]. Artificial intelligence helps humans to do work more efficiently. Artificial intelligence may replace the humans that have weaknesses in many repetitive activities and need rest. Humans have feelings are related to atmospheres and moods that may result in misjudgment and affect a critical decision. The repetitive and monotonous works can be assisted by artificial intelligence [3]. A study reported an expert system knowledge-based approach for Hajj Pilgrims. A dynamic knowledge-based approach was planned to diagnose the expert's possible problems and solutions [4]. Users may ask any questions about the practice of Hajj. Another study in the Islamic domain presented the social network's influence on Islam publishing and serving Islam [5]. Other artificial intelligence, like machine learning, is also widely applied in the Islamic domain.

Machine learning is a part of artificial intelligence that is reliable in facilitating humans to make a decision. Machine learning has a task, namely, classification. Classification thinks faster than humans and can perform multi-tasks [3]. The effect of using the Bayesian Network to the efficacy of the classification model in providing the right advice based on the user's response or answer to the system's question was investigated [1]. Preliminary research was also conducted in the field of Islam to determine an eligible person to receive Zakat [3].Both studies utilize the Bayesian Network (BN) algorithm. Based on the research results obtained that the Bayesian Network algorithm is accurately used for the Islamic domain. Bayesian networks also obtained promise results in an expert system to diagnose social issues based on Quran and Hadith [6].

Even though there have been many studies on the classification in Islamic jurisprudence, but they merely used a rule-based for obtaining a conclusion. Recently, unstructured data in text formed made it challenging to classify. A paradigm shift in data growth has occurred, from mostly organized, not too much, to mostly unstructured [7]. Due to the complicated nature of the text, analyzing, understanding, organizing, and sorting through text data is difficult and time-consuming, so most businesses fail to extract value from it [8]. Text classification is potentially techniques to be used for many domains.

There have been several studies developed using text classification. Multivariate Bernoulli Naïve Bayes Classification and Multinomial Naïve Bayes Classification were used to determining whether the news article's impression was positive or negative [9]. This research also tested the best algorithm.

Based on their results, Multinomial Naïve Bayes Classification is better than Multivariate Bernoulli Naïve Bayes Classification on 312 data. A study was conducted to determine the news topic[10]. The study experiments the proposed model on a real news dataset, and the experiment's result reveals that the modified model is performing reasonably well.

Text classification using machine learning was also conducted in the Islamic domain[11]. The research revealed that text classification based on unstructured data was challenging. Naïve Bayes, SVM, and KNN algorithms were employed in that experiment. Furthermore, a deep learning algorithm was used for Arabic text classification[12]. They reported that the results obtained accuracy 96%.

Based on the literature review, there has been some research on text classification. Two studies were implemented in the Islam domain[11,12]. A deep learning algorithm was used as the best algorithm, but it did not effectively use small data[8]. However, no algorithm works best with the 'No Free Lunch' theorem for every problem[13]. Therefore, this study proposes text classification using some machine learning algorithms, e.g., Bayesian Network and Naïve Bayes, that recognized more accurate, fast, and robust [1,3,6] for Islamic jurisprudence domain.

2. Material and methods

The main source of the data were obtained from a corpus collected from various books[14,15] and website [16] that contain Islamic jurisprudence knowledge. In this study, we limit the source of the bot's knowledge-based to books and websites that are originated from Ustadz Abdul Somad (UAS). UAS is chosen because he is a famous Islamic preacher with millions of followers from Indonesia, Malaysia, Singapore and Brunei, and more important is the abundance of textual resources and questions (text) posted to UAS. A brief introduction, UAS is also known as Dato Seri Ulama Setia Negara, Tuan Guru Abdul Somad, or referred as Syaikh Abdul Somad. He is considered one of the most influential Islamic Preacher in Indonesia, and even the most followed Islamic Scholar in Indonesia [2]. His area of expertise is in Comparative Fiqh.

From the corpus, we have obtained two attributes as part of the initial data. After pre-processing of text, 589 attributes are generated using string to word vector filter. The data type consists of a string and numeric. Figure 1 briefly explains the methodology that has been implemented.

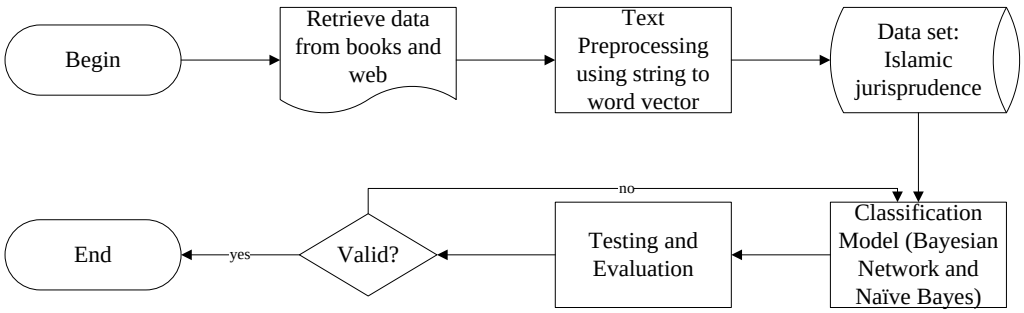


Figure 1. The methodology of text classification based on Islamic jurisprudence using machine-learning techniques.

2.1. Text Pre-processing

To create good quality dataset, we have pre-processed the data. Text Pre-processing involves multiple processes, such as string to a word vector. String to word vector is a filter for converts string attributes from the text contained in the strings into a set of numeric attributes representing word occurrence information. The filter can also operate with a different *stopword* list (*Rainbow* system) than the built-in one. From the first lot of data filtered (training), the dictionary is determined. When a class attribute is set, this filter is not purely unsupervised as it generates a separate dictionary for each class and then

merges it. The snowball stemmer was used for stemming processes. Meanwhile, delimiters (*.,;:'"()?!*) were used for tokenization.

2.2. Classification Model

In this study, we have used two algorithms, Naïve Bayes and Bayesian Network. Let *D* be the random variable that denotes an instance's category and let $X\ (X_1, X_2,...,Xn)$ be a random variables vector that denotes the observed values of the attribute. Let *dj* denote the class tag and let $x\ (x_1, x_2, \ldots, xn)$ represent a particular observed value vector of the category. Bayes theorem was used to calculate the probability as follows to determine the category of sample x:

$$p(D = dj \mid X = x) = \frac{p(D = dj)p(X = X \mid D = dj)}{p(X = x)} \quad (1)$$

Let $U = \{text, jurisprudence\}$ be a set of variables. The Bayesian network *B* is a network structure over a set of variables *U*, which is a Directed Acyclic Graph (DAG) over *U* and a set of tables of probability $B_P = \{p(u|pa(u))|u \in U\}$, which is$p(u)$is the organization of u in structure [11]. A Bayesian network is a distribution of probabilities $P(U) = \Pi_{u \in U} p\left(u|pa(u)\right)$. The classifier has learned from a sample set of data (*attributes, jurisprudence*). The training task is to find a suitable Bayesian network with a collection of data *D* over *U*.

2.3. Testing and Evaluation

For classification modeling, each experiment was executed using the training set, 10-Fold cross-validation. The parameters to test each modeling's output as follows: accuracy, Root Mean Square Error (RMSE), time taken and confusion matrix.

3. Result and Discussion

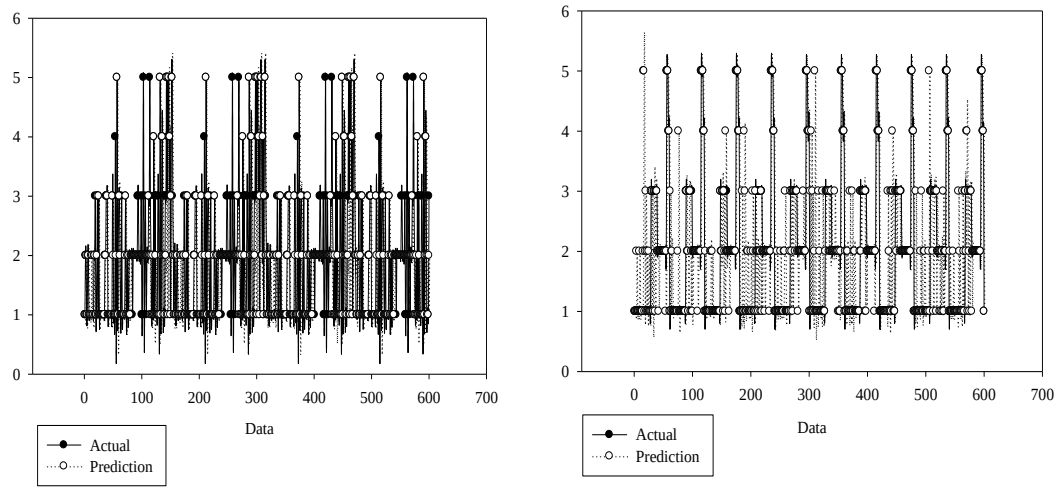
We have successfully used Bayesian Network and Naïve Bayes algorithms to text classification based on Islamic jurisprudence using machine learning techniques. Table 1 shows the comparison of evaluation results for Bayesian Network and Naïve Bayes.

Table 1. The comparison of evaluation results for Bayesian Network and Naïve Bayes.

Evaluation Model	Bayesian Network				Naïve Bayes			
	Accuracy (%)	**RMSE**	**Kappa Statistic**	**Time Taken (second)**	**Accuracy (%)**	**RMSE**	**Kappa Statistic**	**Time Taken (second)**
Use Training Set	74.87	0.26	0.62	0.17	84.25	0.20	0.77	0.27
10-Fold cross-validation	62.98	0.31	0.43	0.12	76.54	0.25	0.66	0.14

Based on the experimental results, the Naïve Bayes algorithm is more accurate in all evaluation models, i.e., 84.25% using training set and 76.54% using 10-Fold cross-validation evaluation models. Likewise, in terms of RMSE, the Naïve Bayes algorithm error rate is lower than the Bayesian Network algorithm. Although the Naïve Bayes algorithm is more accurate but required much time for one execution. Meanwhile, the Bayesian Network algorithm is faster on all evaluation models. Therefore, we concluded that the Naïve Bayes algorithm is more involved in text classification but more accurate than the Bayesian Network.

A detailed comparison between the two algorithms can be ensured in the following Figure 2. Figure 2 shows the comparison between actual and predicted in 600 total data.



(a) *(b)*

Figure 2. The comparison between actual and predicted on total data for *(a)* Bayesian Network and *(b)*Naïve Bayes.

Based on the tests and evaluations, we concluded that the Naïve Bayes algorithm is more accurate than the Bayesian Network algorithm for text classification based on Islamic jurisprudence data. Hence, we can use this Naïve Bayes model to predict the new text data. We have also attempted to make predictions using the new data.

4. Conclusion

We have presented an effective classifier to extract and learn knowledge from Islamic jurisprudence text set using the Bayesian Network algorithm and the Naïve Bayes algorithm. Both algorithms provide excellent accuracy and have a low error rate. By the 'No Free Lunch Theorem,' this study has shown that each algorithm has its own advantages. The Naïve Bayes algorithm has higher accuracy than the Bayesian Network algorithm. However the Bayesian Network has advantages in the time taken, i.e., is faster in run classification tasks. Thus, it can be concluded that the text classification model using Naïve Bayes and String to Word Vector filter have the potential to be used effectively but still has plenty of room for improvement.

This study is still in the early experimental stage. We still need additional experiments such as using Naïve Bayes Multinomial, Naïve Bayes Multinomial Text, Artificial Neural Network, SVM, and other stemming algorithms.

References

[1] Kurniawan, R., Akbarizan, Jamal, K., Nur, A., Ahmad, M. Z., and Kholilah, D., 2018, "Advise-Giving Expert Systems Based on Islamic Jurisprudence for Treating Drugs and Substance Abuse," J. Theor. Appl. Inf. Technol.

[2] Tempo Media, "Lima Ulama Berpengaruh Terhadap Pemilih Versi Survei LSI Denny JA - Nasional Tempo.Co" [Online]. Available: https://nasional.tempo.co/read/1146499/lima-ulama-berpengaruh-terhadap-pemilih-versi-survei-lsi-denny-ja/full&view=ok. [Accessed: 07-Nov-2019].

[3] Akbarizan, Kurniawan, R., Nazri, M. Z. A., Abdullah, S. N. H. S., Murhayati, S., and Nurcahaya, 2019, "Using Bayesian Network for Determining The Recipient of Zakat in

BAZNAS Pekanbaru."

[4] Sulaiman, S., Mohamed, H., Arshad, M. R. M., Rashid, N. A., and Yusof, U. K., 2009, "Hajj-QAES: A Knowledge-Based Expert System to Support Hajj Pilgrims in Decision Making," *ICCTD 2009 - 2009 International Conference on Computer Technology and Development*.

[5] Nassar, I. A., Hayajneh, J. A., and Almsafir, M. K., 2013, "The Influence of Using Social Network on Publishing and Serving Islam: A Case Study of Jordanian Students," *Proceedings - 2012 International Conference on Advanced Computer Science Applications and Technologies, ACSAT 2012*.

[6] Kurniawan, R., Nur, A. M., Yendra, R., and Fudholi, A., 2016, "Prototype Expert System Using Bayesian Network for Diagnose Social Illness," J. Theor. Appl. Inf. Technol.

[7] IBM, "The Biggest Data Challenges That You Might Not Even Know You Have - Watson" [Online]. Available: https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/. [Accessed: 07-Nov-2019].

[8] MonkeyLearn, "Text Classification: A Comprehensive Guide to Classifying Text with Machine Learning" [Online]. Available: https://monkeylearn.com/text-classification/#machine-learning-based-systems. [Accessed: 07-Nov-2019].

[9] Singh, G., Kumar, B., Gaur, L., and Tyagi, A., 2019, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, IEEE, pp. 593–596.

[10] Li, Z., Shang, W., and Yan, M., 2016, "News Text Classification Model Based on Topic Model," *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, IEEE, pp. 1–5.

[11] Rostam, N. A. P., and Malim, N. H. A. H., 2019, "Text Categorisation in Quran and Hadith: Overcoming the Interrelation Challenges Using Machine Learning and Term Weighting," J. King Saud Univ. - Comput. Inf. Sci.

[12] Elnagar, A., Al-Debsi, R., and Einea, O., 2020, "Arabic Text Classification Using Deep Learning Models," Inf. Process. Manag., **57**(1), p. 102121.

[13] Kurniawan, R., Nazri, M. Z. A., Irsyad, M., Yendra, R., and Aklima, A., 2015, "On Machine Learning Technique Selection for Classification," *Proceedings - 5th International Conference on Electrical Engineering and Informatics: Bridging the Knowledge between Academic, Industry, and Community, ICEEI 2015*.

[14] Abdul Somad, 2013, *77 Tanya Jawab Seputar Shalat: Shalatlah Sebagaimana Kalian Melihatku Shalat*, Tafaqquh Study Club, Pekanbaru.

[15] Abdul Somad, 2015, *37 Masalah Populer: Untuk Ukhuwah Islamiyah (37 Masalah Populer)*, Tafaqquh, Pekanbaru.

[16] Abdul Somad, 2019, "Somadmorocco.Blogspot.Com" [Online]. Available: http://somadmorocco.blogspot.com/. [Accessed: 07-Nov-2019].